Manipuler les données. Documenter le marché

Jean-Sebastien Vayre

MANIPULER LES DONNÉES. DOCUMENTER LE MARCHÉ

Les implications organisationnelles du mouvement big data

JEAN-SEBASTIEN VAYRE

Qu'est-ce que le big data ? Comment le caractériser ? Quel est son impact sur l'organisation du marché ? D'un point de vue marchand, le big data consiste à transformer les traces d'activités des consommateurs en informations dans le but de documenter les acteurs du marché. Le big data est donc un processus de documentation. D'abord, nous soutenons que la dynamique de ce processus renvoie à une évolution sociotechnique et une révolution sociocognitive. Ensuite, nous pointons les implications de cette révolution sur l'organisation du marché. En conclusion, nous soulignons que si le processus de documentation big data peut favoriser la réactivité et l'adaptabilité des organisations marchandes, il peut également engendrer des biais de connaissance importants sur le plan socioéconomique.

Introduction

Au début du XX$_{ème}$ siècle, les publicités ont pour rôle de mettre en scène les aspects utilitaristes des produits (Courbet, 2006). Elles témoignent ainsi d'un rapport particulier à l'information. À cette époque, l'information marchande est en effet définie comme parfaitement objective. C'est pourquoi, les spécialistes du marché la diffusent dans le but d'équiper la rationalité substantive des consommateurs. Ici, le problème de l'information marchande est donc saisi à travers la figure de l'*homo oeconomicus*. C'est-à-dire, un consommateur capable de classer ses préférences selon les principes de la transitivité et de mobiliser son environnement afin de maximiser satisfaction et utilité (Laval, 2007). Par conséquent, à cette époque, le problème de l'information marchande renvoie directement à celui du calcul d'un consommateur considéré comme universel.

Dès la deuxième moitié du XX$_{ème}$ siècle, les formes de la communication publicitaire sont profondément modifiées. À ce moment, les publicités reflètent moins une argumentation articulant logiquement des informations dites utiles qu'une logique d'association de symboles et de connotations destinés à flatter les désirs profonds des consommateurs (Barthes, 1964 ; Ditcher, 1961). Partant, les marketers ne définissent plus l'information marchande comme objective et universelle. Ils considèrent plutôt que c'est l'environnement social dans lequel baignent les consommateurs qui fournit le système de règles permettant d'interpréter l'information (Gomez, 2006). Dès lors, les documents publicitaires ont pour finalité d'équiper la rationalité située et limitée des consommateurs (Simon, 1982). Cette fois-ci, la question de l'information marchande est donc moins saisie à travers la figure de l'*homo oeconomicus* que par le biais de celle de l'homme psychosociologique. C'est-à-dire, un consommateur dont les raisonnements sont sous-tendus par des heuristiques (i.e. : raccourcis mentaux) qui sont largement socialisées (Tversky et Kahneman, 1974). Aussi, jusqu'au début du XXI$_{ème}$ siècle, l'information publicitaire est constituée selon un double processus de compréhension et de séduction : elle est directement associée au problème de la captation d'un groupe de consommateurs (Cochoy, 2004).

Actuellement, le problème de l'information marchande connaît un autre changement important. Il devient une sorte de synthèse augmentée des deux problématiques précédentes. En effet, les documents publicitaires d'aujourd'hui ont globalement pour objectif d'équiper la rationalité substantive des consommateurs, de capter leurs désirs profonds, mais aussi et surtout d'« économiciser » une partie de leur cognition (Kessous, 2012). Car, avec le développement des Technologies Numériques de l'Information et de la

Communication (TNIC)[1], les documents publicitaires finissent par intégrer un processus consistant à transformer les traces d'attention des consommateurs en informations susceptibles de les intéresser. Et, dans cette économie bien particulière, la question de l'information marchande est appréhendée à travers la figure de l'homme cognitif. C'est-à-dire, un consommateur qui infère les informations de son environnement selon des contraintes sociales, relationnelles, mais aussi biologiques et fonctionnelles qui régulent et conditionnent ses activités mentales (Weil-Barais, 2005). De ce fait, que ce soit sur le plan purement cognitif, c'est-à-dire du traitement de l'information, ou relationnel, c'est-à-dire des intérêts et des passions, l'information marchande renvoie aujourd'hui au problème de l'attention d'un consommateur bien singulier.

Dans un contexte de consommation de masse et de numérisation du marché où les mots d'ordres sont personnalisation, pertinence et précision, ce problème d'information-attention prend alors tout son intérêt. Car, il apparaît maintenant qu'une des plus importantes difficultés pour les professionnels du marché est de trouver les moyens de communication, ou si l'on préfère, les systèmes d'information qui leurs permettent d'interagir efficacement et rapidement avec une masse de consommateurs considérés dans leurs singulières pluralités.

À en croire les spécialistes de la donnée intéressés par le marché, il existe maintenant une solution à ce problème. La solution, c'est le big data ! L'argument est le suivant. Les consommateurs d'aujourd'hui mobilisent activement les technologies numériques pour effectuer leurs activités de consommation. Ils produisent ainsi une masse considérable de données qui peuvent permettre aux marketers de comprendre et prédire plus finement leurs comportements. En ce sens, ces données sont susceptibles d'aider les marketers à résoudre le problème de l'information-attention. Mais, pour cela, les marketers doivent avant tout être capable d'emmagasiner et de calculer cette

[1]1. Nous utilisons le sigle TNIC afin de souligner l'évolution générale des Technologies de l'Information et de la Communication. Car, dès les années 1970, les chercheurs et les spécialistes utilisent le concept de Nouvelles Technologies de l'Information et de la Communication (NTIC) afin de désigner les dernières évolutions de la télématique. À partir des années 2000, du point de vue de la littérature scientifique, le N des NTIC tend lentement à disparaître : les Technologies de l'Information et de la Communication n'étant plus vraiment Nouvelles, les spécialistes préfèrent parler des TIC. Maintenant, le N réapparaît progressivement ; mais, à un autre emplacement. Il correspond au terme « numérique » et permet ainsi d'intégrer les évolutions actuelles comme, par exemple, le développement des objets communicants, des systèmes embarqués, des dispositifs intelligents, de l'informatique en nuage, etc.

grande quantité de données. Or, le big data a précisément pour fonction de stocker et traiter de très gros volumes de données.

Partant, comme toutes les solutions qui apparaissent trop évidentes, le big data a quelque chose de mystérieux. Quelle est cette sorte de technologie miracle qui pourrait permettre de lever un des plus importants défis du commerce d'aujourd'hui ? Autrement dit, qu'est-ce que c'est que le big data ? D'où vient-il et comment a-t-il été développé ? Est-ce une simple évolution ou une véritable révolution ? Et, finalement, quelle est son influence sur l'organisation du marché ?

Pour répondre à ce questionnement, nous proposons, dans la première section, de définir le big data comme un processus de documentation. Nous montrons alors que ce processus peut être saisi selon deux niveaux d'observation. L'observation profonde permet de focaliser sur la dimension cognitive du big data ; et, celle de surface, sur sa dimension technique. Nous parlons alors de système d'information. Dans la deuxième section, nous proposons d'examiner le big data du point de vue de l'observation de surface. Nous soutenons ici que le système d'information big data est avant tout une évolution sociotechnique. Le big data recouvre un complexe d'innovations qui, largement influencées par le marché, intègrent la dynamique générale des TNIC. Dans la troisième section, nous étudions le big data selon le point de vue dit profond. Nous soutenons cette fois-ci que, dans le domaine du marché, le processus de documentation big data constitue une révolution sociocognitive. Car, à partir des technologies d'apprentissage artificiel, les marketers les mieux équipés sont capables d'extraire des connaissances de façon tout à fait nouvelle. Dans la quatrième section, nous discutons les implications organisationnelles qui sont associées à cette révolution. Nous concluons en montrant que le processus de documentation big data tend à favoriser un mode d'organisation du marché plus hétérarchique.

1. Un processus de documentation

Nous l'avons vu, au regard des spécialistes de la donnée intéressés par le marché, le big data doit permettre de résoudre le problème de l'information-attention. Et, comme nous venons de l'exposer, nous ne proposons pas, dans cet article, d'expliquer pourquoi ces spécialistes auraient tort ou raison, mais plutôt de comprendre comment le big data a progressivement été développé et comment certaines de ses applications peuvent impacter l'organisation du marché. Toutefois, avant d'examiner ces deux phénomènes, nous souhaitons commencer par définir ce qu'est le big data.

Big data est un terme à la mode qui est issu des sciences de la gestion. Cette notion a émergé dans les bureaux de la Silicon Valley des années 1990. Cependant, le concept de big data a pris son véritable essor avec la diffusion des trois principaux rapports que l'institut Mc Kinsey a rédigé sur le sujet : *Clouds, big data, and smart assets : Ten tech-enabled business trends to watch* (Bughin *et al.*, 2010), *Are you ready for the era of « big data »* (Brown *et al.*, 2011) et le classique *Big data : The next frontier for innovation, competition and productivity* (Manyika *et al.*, 2011).

Dans cette section, nous proposons de saisir le big data comme un processus de documentation. Nous commençons par dégager la double définition que sous-tend le concept de « big data » (1.1). À partir de cette double définition, nous montrons que le big data constitue un processus de documentation qui peut être compris selon deux niveaux d'observation. Le premier niveau (1.2) doit permettre au chercheur de saisir ce qui se joue sur le versant cognitif du big data. Le deuxième niveau doit permettre au chercheur de comprendre ce qui se trame sur le versant technique du big data. Nous concluons cette première section en présentant le corpus de données que nous avons mobilisé pour étudier la dynamique du processus de documentation big data (1.3).

1.1. Une double définition

La notion de « big data » est issue du monde anglo-saxon. Elle est généralement traduite par l'expression « grosses données » ou « données massives ». Selon cette première définition, du point de vue marchand, les big data désignent cette gigantesque quantité de données quotidiennement produites à partir des interactions que les clients ont avec leur environnement numérique. Cependant, il apparaît que les acteurs intéressés par ces données préfèrent parler du big data que des big data. Il découle de ce glissement vers le singulier une seconde définition plus générale. Ici, le big data, c'est ce mouvement qui consiste à développer des systèmes de production et de calcul de données numériques en intégrant les technologies de télécommunications, de stockage et traitement des données, et de l'intelligence artificielle.

En outre, bien que le référent du terme « big data » soit plus ou moins extensif, son objectif n'en reste pas moins clairement circonscrit. Au regard de l'ensemble des acteurs des sphères économique, politique, scientifique et technique, le big data/les big data a/ont pour principale finalité de produire des connaissances plus objectives. C'est-à-dire, dans le cas du marché, d'opérer des combinaisons de données factuelles capables d'informer rigoureusement les acteurs de l'offre et de la demande. C'est en ce sens que le big data marchand

constitue un processus de documentation. Et, en référence à la double définition que nous venons d'exposer, ce processus peut être appréhendé selon deux axes d'observation.

1.2. De l'observation profonde à l'observation de surface

Le premier axe, que nous qualifions de profond, se joue sur un plan cognitif. Il consiste à suivre les processus de re-présentation qui vont du traçage des interactions Homme-Machine à leur transformation en documentations (Latour, 1993). Plus concrètement, à travers l'observation profonde, le chercheur doit pouvoir repérer l'enchaînement des transcriptions qui permettent aux interactions Homme-Machine de devenir des données non-structurées (e.g. : un texte, une image, un son, etc.), ou/puis, des données structurées (e.g. : une unité linguistique, un nombre de clic, une coordonnées GPS, etc.), pour enfin passer à un état d'information (e.g. : un pourcentage, une corrélation, une prédiction, etc.), et, de documentation (e.g. : une analyse de requête, un rapport de géolocalisation, de navigation, etc.).

Le second axe, que nous qualifions de surface, se joue, quant à lui, sur un plan technique. Ici, il s'agit de suivre l'évolution du complexe technologique que compose le processus de documentation big data. L'observation de surface doit permettre au chercheur de dégager la dynamique d'intrication des sept groupes d'éléments suivants. Le premier est celui des transcripteurs qui assurent l'encodage des interactions Homme-Machine. En d'autres termes, ce sont eux qui prennent en charge l'attachement de l'émetteur et la traduction de son action en une donnée (non-)structurée (e.g. : une montre, une tablette, un téléphone, etc., équipé(e) d'un logiciel de capture adapté). Ensuite, les algorithmes viennent, en quelque sorte, prendre le relais des transcripteurs. Ils ont en effet pour principale fonction de traiter et assembler les données (non-)structurées afin de produire une ou plusieurs information(s). Alors, les montreurs prolongent le travail des algorithmes : ils présentent les informations selon diverses techniques de visualisation pour constituer un ou plusieurs document(s). Et, finalement, les transmetteurs ont pour fonction de garantir le décodage des documents ; c'est-à-dire, l'attachement du récepteur et la transmission des données (e.g. : une montre, une tablette, un téléphone, etc., équipé(e) d'un logiciel de traitement adapté). Ensuite, les trois derniers groupes sont ceux des réseaux de télécommunications, des centres de données et de l'informatique en nuage qui ont respectivement pour fonction d'assurer la circulation et la conformité des données dans l'espace, l'organisation et la préservation des données dans le temps, et de garantir la bonne intégration de chaque groupe d'éléments.

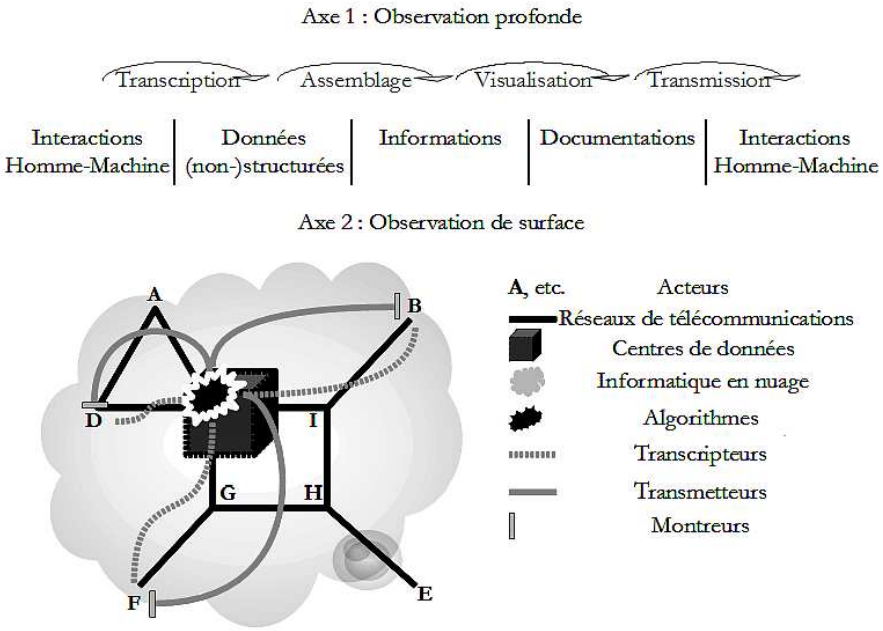


Figure 1 : Le processus de documentation big data

1.3. Corpus de données

Afin d'étudier le processus de documentation big data selon les deux axes d'observation que nous venons d'exposer, nous avons mobilisé un corpus composé de 188 documents de seconde main.

Dans la section 2, nous mobilisons l'analyse qualitative et quantitative de 142 documents d'archive du *New York Times*. Nous avons sélectionné ces documents à l'aide du mot-clé « big data » et des filtres d'ancienneté et pertinence. Notons alors que, bien que l'objectivité de ce corpus soit discutable (Bourdieu, 1996), il constitue néanmoins un ensemble de données pertinentes. Car, chaque archive, en fixant le point de vue d'un professionnel du journalisme, porte la trace d'un angle de vue particulier : celui d'un acteur social qui vit de l'information. En elle-même, une de ces archives présente donc une expertise bien singulière. Cependant, une fois associée aux autres, l'ensemble offre un point d'observation plus général qui permet d'explorer rigoureusement le phénomène représenté.

Dans les sections 3 et 4, nous nous appuyons principalement sur l'analyse qualitative de vingt-huit articles de journaux et de dix-huit courts entretiens vidéo. Les articles de journaux ont été écrits par des professionnels du big data et diffusés dans les bulletins du Salon du Big Data 2013. Les courts entretiens vidéo ont également été réalisés auprès de professionnels du big data. Nous les avons sélectionnés à partir du moteur de recherche Google et du mot-clé « big data ». Nous les avons retranscrits dans leur intégralité. Comme dans le cas des archives du *New York Times*, ces documents comportent de nombreux biais. Ils ne sont pas le produit d'un dispositif méthodologique que nous avons pu contrôler et renvoient, dans la plupart des cas, à des objectifs de communication grand public. Toutefois, étant à chaque fois rédigés ou énoncés par des professionnels engagés dans le développement du big data, ces documents témoignent de représentations et pratiques qui sont ancrées dans des expériences de terrain. À la manière des archives du *New York Times*, une fois considéré dans son ensemble, ce corpus permet alors de saisir correctement les enjeux et usages professionnels du big data.

2. Une évolution sociotechnique

Nous venons de présenter la double définition du big data. Dès lors, nous souhaitons dégager les formes de son développement. Car, nous avons vu que le big data forme un processus de documentation et que ce processus est sous-tendu par des mécanismes de production, stockage et traitement d'information qui sont organisés les uns par rapport aux autres. Ces mécanismes renvoient donc à une architecture technologique bien spécifique. Et, c'est justement l'évolution de cette architecture que permet de mieux comprendre l'observation surface.

Ainsi, à partir des catégories d'analyse que nous avons dégagé dans la section 1.2, nous soutenons que, d'un point de vue sociotechnique, le système d'information big data relève plutôt du domaine de l'innovation incrémentale que de celui de l'innovation radicale. Pour ce faire, nous présentons la conception des quatre graphiques sur lesquels repose notre développement (2.1). Nous commençons alors par discuter la dynamique technologique du big data (2.2) ; puis, sa dynamique sociale (2.3). Nous finissons en montrant que le développement du système d'information big data est avant tout de l'ordre de l'évolution sociotechnique (2.4).

2.1. Présentation

Les deux figures subséquentes représentent, du point de vue des 142 archives du *New York Times*, l'évolution technique (cf figure 2) et sociale (cf figure 3) du big data. Elles ont été conçues de la façon suivante. Nous avons effectué plusieurs lectures des 142 documents d'archive et nous avons dégagé quatorze catégories d'analyse. Sur ces quatorze catégories, nous avons sélectionnés les huit les plus significatives :

- celle des réseaux de télécommunications ;
- celle de l'informatique en nuage ;
- celle des centres de données ;
- celle des transcripteurs et des transmetteurs ;
- celle des algorithmes ;
- celle de la protection de la vie privée ;
- celle des définitions et états des lieux du big data ;
- et, celle des critiques du big data.

Afin de faciliter la lecture des quatre graphiques, nous avons construit les huit périodes qui sont représentées sur l'axe des abscisses. La première s'étend de 1991 à 2001 et la seconde de 2002 à 2007. Les six autres périodes sont, à chaque fois, une année révolue. Nous avons choisi cette échelle irrégulière afin de pouvoir disposer d'effectifs totaux relativement comparable pour chaque période. Car, nous avons recueilli 12 documents d'archive sur la période 1991-2001, 12 sur celle de 2002-2007, 17 sur celle de 2008, 13 sur celle de 2009, 18 sur celle de 2010, 28 sur celle de 2011, 21 sur celle de 2012 et 21 sur celle de 2013. Or, afin de permettre aux lecteurs de mesurer facilement le poids de chaque catégorie dans une période donnée, nous avons souhaité présenter des pourcentages sur l'axe des ordonnées. Nous avons donc préféré privilégier la relative comparabilité des effectifs totaux à celle des durées de chaque période. À titre d'exemple, le graphique 1 de la figure 2 présentée ci-dessous indique ainsi qu'il y a 40% des 28 documents d'archives recueillis sur la période 2011 qui traitent de la thématique de l'informatique en nuage[2].

[2]2. Rappelons que les effectifs totaux d'une période comprennent l'ensemble des documents d'archive toutes catégories confondues ; c'est-à-dire, les quatorze catégories dans leur ensemble et non pas seulement les huit que nous avons sélectionné ici.

2.2. Une dynamique technologique...

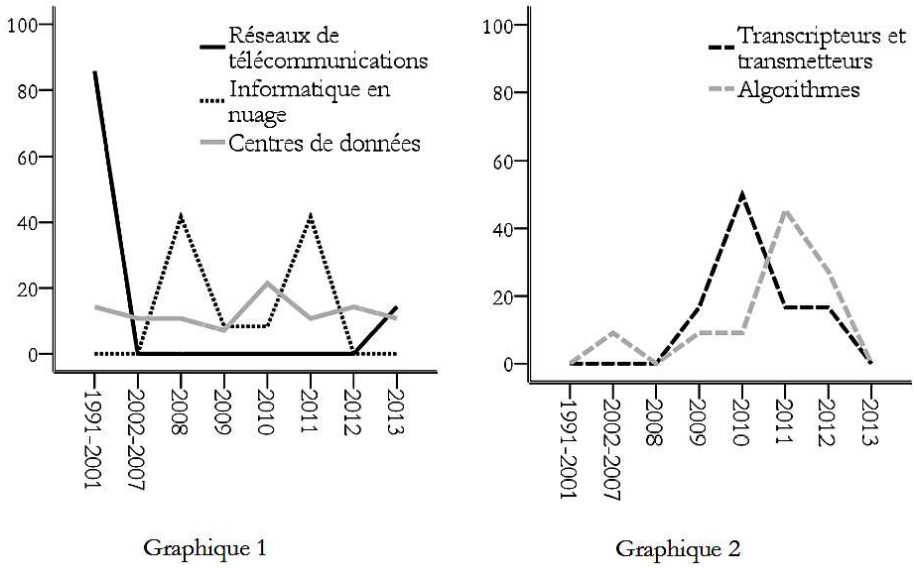


Figure 2 : Dynamique technologique du big data

À partir des graphiques 1 et 2, il est possible de dégager quatre grandes périodes. La première, qui s'étend de 1991 à 2001, marque le développement des réseaux de télécommunications. La seconde, qui va de 2002 à 2008, représente l'émergence des technologies de l'informatique en nuage. La troisième, qui se déroule de 2009 à 2012, montre la dernière phase d'évolution du big data. Pour finir, la période actuelle (i.e. : 2013) est marquée par la décroissance des problématiques technologiques.

Plus précisément, l'analyse qualitative du corpus montre qu'entre 1991 et 2001, le big data est au tout premier stade de son développement. À ce moment, les principales évolutions technologiques associées au big data se produisent dans le domaine des réseaux de télécommunications et des centres de données. Les usages que les journalistes du *New York Times* font du terme big data témoignent de cet état liminaire. Il est avant tout mobilisé pour qualifier des acteurs comme Ascend Communication ou encore Oracle qui sont respectivement spécialisés dans les télécommunications et le stockage et traitement des données. Partant, en 1998, Bell Atlantic met en place le réseau de télécommunications longues distances aux États-Unis. Dans le même temps,

American Telephon & Telegraph délègue progressivement ses activités de gestion de la relation client à International Business Machines. Alors, d'autres acteurs, comme Castle Networks, Argon Networks, Siemens, Alcatel, Xylan et Computer Communication Compatibility entre en scène. Les stratégies de compétition et coopétition se multiplient. Il s'instaure ainsi un double mouvement de concurrence économique et de spécialisation technologique qui a pour conséquence de promouvoir et démocratiser les « Nouvelles Technologies de l'Information et de la Communication ». D'après la banque mondiale, il y a aux États-Unis 140 147 752 internautes en 2001 contre 2 975 535 en 1991 ; 12 792 812 abonnés à l'Internet haut débit en 2001 contre 705 000 en 1998 ; et, 128 500 000 abonnés de téléphone mobile en 2001 contre 7 557 148 en 1991[3].

Par conséquent, en 2002, après l'éclatement de la bulle Internet de 2001, une bonne partie des collectivités, des entreprises et des citoyens américains sont connectés à la toile. Le problème de la circulation et du stockage des données finit alors par prendre le pas sur celui du développement des réseaux de télécommunications. Dans la société de l'information qui est en train de se constituer, les données numériques occupent en effet une place toujours plus remarquable. Elles doivent circuler de la façon la plus fluide possible ; et surtout, elles doivent être sauvegardées pour pouvoir exister ! Les acteurs que sont Google, Yahoo, Microsoft, Facebook, Amazon, Egan Marino Corporation et Hewlett-Packard relancent alors la course économique à l'innovation technologique. Selon un double processus de concurrence et de spécialisation analogue à celui présenté en amont, ce sont cette fois-ci les technologies de stockage et traitement des données qui commencent à se démocratiser. C'est ainsi que dès 2008, les grands acteurs du stockage et du traitement des données décident de valoriser leurs compétences en proposant différents services d'informatique à la demande : c'est le temps de l'informatique en nuage. Une technologie qui a alors pour fonction de fluidifier la circulation des informations en intégrant l'ensemble des terminaux numériques au sein d'une puissance de stockage et traitement des données toujours plus colossale.

Cependant, suite à la crise financière de 2008, il semble que le phénomène big data entre dans une sorte de renfoncement technologique plus ou moins associé au mouvement de dépression économique. En tout cas, dès 2010 émerge un nouvel intérêt pour les problématiques de stockage et de traitement des données. Ce regain d'attention est alors doublé par l'avènement des technologies de transcription/transmission. Par exemple, des acteurs comme

[3]3. Ces chiffres sont diffusés sur le Journal du Net. Ils sont consultables à l'adresse suivante : http://www.journaldunet.com/web-tech/chiffres-internet/etats-unis/pays-usa.

Body Média commercialisent des systèmes de captation capables d'enregistrer plus de 9000 variables physicobiologiques[4]. Compte tenu du fait que le développement croissant de ces systèmes est étroitement couplé à ceux du web sémantique et de l'Internet des objets, les centres de données sont conduits à stocker en continu des masses de données à la fois plus diversifiées et moins structurées. Comme l'indique la série de concours organisés par Netflix, les innovations dans le domaine du traitement des données sont ainsi placées au premier plan. En d'autres termes, et nous y reviendrons dans la troisième section, à travers les algorithmes de classification et de prédiction, l'apprentissage artificiel (i.e. : machine learning) devient le centre de gravité du système d'information big data.

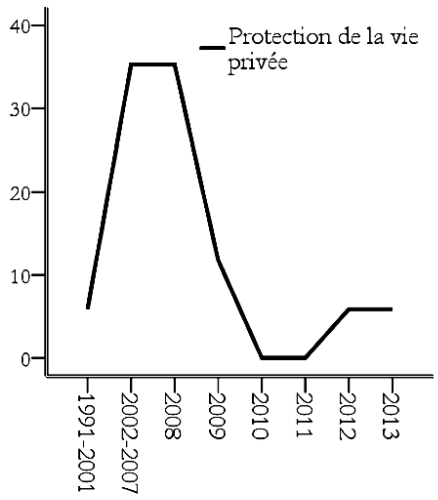
Par conséquent, bien que le problème de visualisation n'apparaisse pas clairement dans les archives du *New York Times*, le système d'information big data est, en 2012, parfaitement mature sur le plan technique. Et, comme nous allons le voir, en 2013, il fait l'objet de nombreuses critiques.

2.3. … ou une dynamique sociale ?

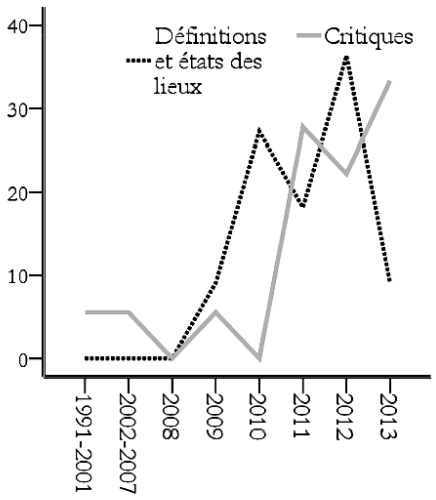
Ces critiques indiquent bien que le big data ne peut être correctement saisi à partir de sa seule dynamique technologique. En effet, il est animé par des controverses où sont confrontés des imaginaires, des convictions, des expériences, etc., qui viennent façonner son développement.

---

[4]4. Plus précisément, il s'agit de capteurs brassards qui peuvent enregistrer les activités physiques, le nombre de calories brûlées, la chaleur du corps, l'efficacité du sommeil, etc.



Graphique 3



Graphique 4

Figure 3 : Dynamique sociale du big data

Dans ces graphiques, nous pouvons repérer deux grandes périodes. La première, qui s'étend de 2002 à 2009, montre l'étendue du problème de la protection de la vie privée. La seconde, qui va de 2010 à 2013, représente un processus plus général de réflexivité critique.

Dans les archives du *New York Times*, le problème de la protection de la vie privée s'impose à partir de 2002 ; soit, juste après le mouvement de démocratisation de l'Internet décrit précédemment. Plus exactement, il ressort jusqu'à 2007 que la protection de la vie privée est directement associée au problème de l'usurpation d'identité. C'est-à-dire, à la possibilité qu'une personne puisse récupérer les informations privées d'une autre à partir d'une fausse identité. Dans un premier temps, le problème des données personnelles renvoie donc aux questions de la sécurisation des canaux, de l'authentification ou encore du chiffrement à haute vitesse. Par exemple, la sécurité des paiements sur Internet constitue, à ce moment, un important problème de protection des données personnelles. Deux types de positionnements permettent alors de comprendre les enjeux économiques et politiques liés à ces questions d'usurpation d'identité. Dans le premier, certains acteurs, dont font généralement partie les associations de consommateurs, vont prôner un renforcement de la régulation politique de la circulation des données personnelles. Ils soulignent les usages non-contrôlés, voire déviants, qu'il est

possible de faire de ces données. L'objectif sous-jacent est de favoriser la sécurité des consommateurs, même si c'est au détriment de leur liberté. Dans le second, d'autres acteurs, dont font souvent partie les marketers, vont préconiser de laisser le marché équilibrer la circulation des données personnelles. L'objectif sous-jacent est, cette fois-ci, de favoriser la liberté des consommateurs, même si cela se fait au détriment de leur sécurité.

À partir de 2008, les partisans de la régulation soulignent les difficultés rencontrées pour légiférer efficacement sur la circulation des données à caractère personnel. Mais, surtout, ils pointent l'autre face du problème de la protection de la vie privée : celui de la ré-identification. C'est-à-dire, la possibilité qu'une personne puisse retrouver l'identité d'une autre à partir d'un certain nombre de données pourtant préalablement anonymisées. Ici, ce sont les usages que les marketers font des données de leurs clients qui sont principalement visés. Pour les partisans de la régulation, l'enjeu est alors de permettre à chaque consommateur de pouvoir contrôler l'extension d'identité qu'est son empreinte numérique.

Dès 2010, le problème de la protection de la vie privée finit alors par recouper une réflexion plus générale sur les enjeux des grosses données. Comme si la multiplication des applications et les usages du big data conduisaient les acteurs à éprouver le besoin de (se) renseigner (sur) ses origines et ses évolutions. Il s'agit alors de définir ce qu'est le big data et de réaliser une sorte d'état des lieux de son développement. Les archives du *New York Times* montrent ainsi que le terme « big data » est pour la première fois utilisé dans une série de diapositives réalisées par John Mashey (1998), un ancien expert de la Silicon Graphics, puis, est stabilisé par Francis X. Diebold (2003), un économiste de l'université de Pensylvanie et enfin est institutionnalisé par Randal E. Bryant, Randy H. Katz et Edward D. Lazowska (2008), trois chercheurs en informatique associés au Computing Community Consortium. En parallèle, dès 2011, émerge une série de critiques concernant les implications générales du mouvement big data. Ce sont principalement les qualités objectives et éthiques de ce système d'information qui sont visées. Certains nuancent ainsi la valeur épistémique des connaissances produites par le processus de documentation big data pendant que d'autres pointent les complexités éthiques associées aux enjeux de contrôle qui lui sont sous-jacents. De ce fait, en 2013, suite à un long mouvement de promotion et démocratisation du big data, advient un véritable effort de réflexivité critique.

2.4. Un système sociotechnique en évolution

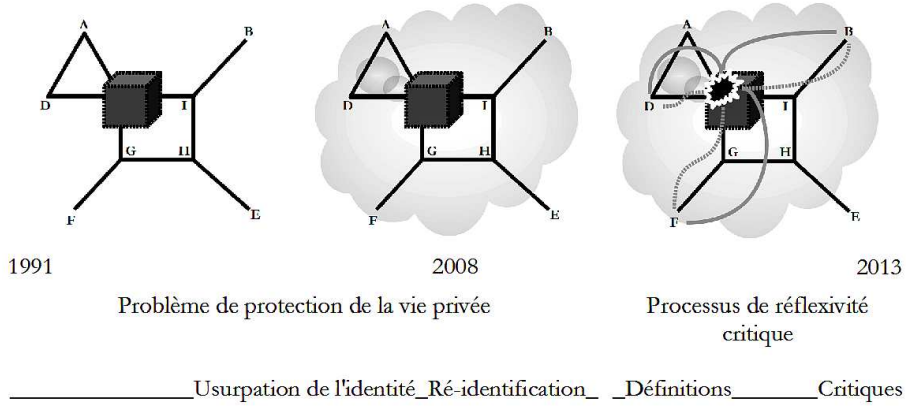


Figure 4 : Dynamique sociotechnique du big data

Comme le schématise la figure 4, le système d'information big data n'est donc ni complètement déterminé sur le plan social, ni complètement déterminé sur le plan technique. Le social et le technique forment en effet une sorte de tissu sans couture tressé durant de multiples négociations entre des acteurs et des actants qui animent les sphères technique, scientifique, politique, économique, juridique, environnementale, culturelle, sociale et humaine qui composent tout mouvement d'innovation (Akrich, 1989 ; Hughes, 1987). Le big data est ainsi un système sociotechnique dont l'économicisation (Callon, 1986) repose sur l'intrication équilibrée (sur le plan économique) et ordonnée (sur le plan sociologique) de neuf espaces d'intéressement et d'entre-définition. Les deux premiers sont ceux de la science et de la technique qui traitent des problèmes de la valeur épistémique et du stockage/traitement des données. Dans les espaces du politique, du droit, de l'économique et de l'écologique sont alors respectivement mis en discussion les problèmes de l'éthique à adopter face au pouvoir de contrôle des données, de la régulation de leur circulation, de la création de leur valeur ajoutée, et de la consommation énergétique des centres qui les stockent et les traitent. Les trois derniers espaces sont ceux de la culture, du social et de l'humain où sont débattues les questions de l'universalité d'une communauté organisée autours des données, des inégalités d'accès et de maîtrise associées à leur production et à leurs usages, et de l'attention cognitive et sociale qu'elles favorisent ou limitent.

En conclusion, du point de vue de l'observation de surface, l'évolution du complexe sociotechnique que compose le système d'information big data relève davantage de l'innovation incrémentale que de l'innovation radicale (Freeman, 1994). Car, comme nous avons pu le remarquer, cette évolution se manifeste plus sous la forme d'une continuité intégrer à la dynamique des TNIC et sous la forme d'un mouvement influencé par le marché et les enjeux d'organisation qui y sont associés, que sous celle d'une véritable rupture sur le plan social et technique.

3. Une révolution sociocognitive

Cependant, du point de vue de l'observation profonde, il en va différemment. En effet, et nous l'avons signalé dans la section 2.2, avec le développement des technologies de transcription/transmission, les dispositifs d'assemblage des données sont placés au cœur du processus de documentation big data. C'est pourquoi, dans le cas du marché, les algorithmes de classification et de prédiction deviennent le centre de gravité. Car, compte tenu de la grande quantité, diversité et fluidité des grosses données, le cerveau humain, même assisté par les outils classiques de fouille de données (i.e. : data mining), ne peut pas traiter efficacement les big data. Les machines doivent donc apprendre à organiser ces données et à dégager des combinaisons pertinentes afin de permettre à l'homme de les explorer et de les interpréter en temps réel.

Dès lors, nous soutenons dans cette section qu'à travers les technologies d'apprentissage artificiel, le processus de documentation big data constitue, du point de vue marchand, une innovation radicale sur le plan sociocognitif. Nous commençons par présenter les trois grandes méthodes d'apprentissage machine : l'apprentissage supervisé, non-supervisé et semi-supervisé (3.1). Ensuite, nous exposons les deux principaux modèles d'apprentissage artificiel : le modèle cognitiviste (3.2) et le modèle connexionniste (3.3)[5]. Nous dégageons les avantages et inconvénients des applications techniques associées à chaque modèle. En conclusion (3.4), nous montrons que les applications marchandes

[5]5. En réalité, il existe trois modèles d'apprentissage artificiel : le modèle cognitiviste, le modèle connexionniste et le modèle statistique. Néanmoins, sur le plan fondamental, les différentes techniques d'apprentissage statistique sont largement influencées par les modèles cognitiviste et connexionniste. Par exemple, la notion d'espace des hypothèses est issue de l'héritage cognitiviste alors que celles de réseau de neurones, de carte de Kohonen ou encore de réseau bayésien, sont, à différents degrés, plutôt issues de l'héritage connexionniste (Cornuéjols et Miclet, 2013). Aussi, afin de permettre aux lecteurs de saisir correctement les enjeux fondamentaux de l'apprentissage machine, nous avons décidé de ne présenter que les seuls modèles cognitiviste et connexionniste.

des technologies d'apprentissage non-supervisé et d'influence connexionniste permettent aux machines actuelles les plus performantes d'exploiter de nouvelles logiques d'exploration des données.

3.1. L'apprentissage artificiel

L'apprentissage artificiel consiste à faire produire par les machines un ensemble de connaissances à partir de données symboliques ou numériques. Les spécialistes de l'intelligence artificielle proposent souvent une double définition de l'apprentissage. Dans la première, l'apprentissage consiste à améliorer la performance d'un agent cognitif dans un domaine d'activité bien circonscrit (Chomsky, 1965). Par exemple, à force de calculer la différence entre deux prix, l'apprenant perfectionne ses aptitudes à appliquer la règle de soustraction. C'est l'apprentissage par amélioration du comportement. Dans la deuxième, l'apprentissage consiste à généraliser la compétence d'un agent cognitif d'une activité à une autre (Chomsky, 1965). Par exemple, si l'apprenant est capable de calculer la différence entre deux prix, il doit également être capable de calculer la différence entre deux poids. C'est l'apprentissage par généralisation.

Dans le domaine de l'intelligence artificielle, ces deux familles d'apprentissage sont réalisées selon trois types de méthodes. La première est celle de l'apprentissage supervisé. Prenons l'exemple de deux vendeurs de téléphone portable : Cornélius (qui joue le rôle de l'agent cognitif artificiel) et Zira (qui joue le rôle du programmeur-analyste)[6]. Zira est une experte de la relation client. Cornélius est, quant à lui, débutant. Aussi, si Zira n'éprouve aucune difficulté à identifier les clients fidèles (F) des infidèles (I), Cornélius en est encore à quantifier les caractéristiques qui lui semblent importantes pour opérer cette distinction. Après plusieurs brouillons, Zira lui propose de retenir les deux indicateurs que son le nombre d'achats et de visites. Cornélius finit alors par se représenter le problème à l'aide du graphique de gauche présenté ci-dessous.

[6]6. Nous avons inventé cet exemple à partir d'une série d'illustrations proposées par Antoine Cornuéjols et Laurent Miclet (2013, pp. xviii-xix). Précisons alors que cet exemple, étant très réducteur, ne rend pas compte de la complexité et de la diversité des formes d'apprentissage artificiel. Néanmoins, il permet de comprendre convenablement les principes élémentaires de l'apprentissage automatique.

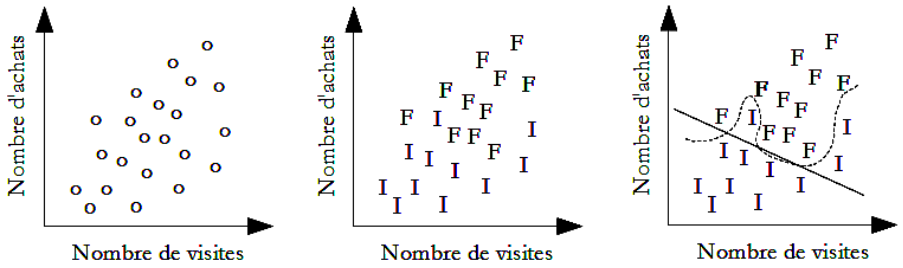


Figure 3 : L'apprentissage artificiel de Cornelius

Ne sachant pas trop quoi faire de son graphique, Cornélius est conduit à se placer dans une relation d'instruction en demandant à Zira de bien vouloir lui souffler les réponses correctes. Celle-ci ayant aimablement accepté, Cornélius s'empresse de dessiner le graphique du centre. Il pose alors le problème d'apprentissage suivant : quelle est la règle qui permet, avec le plus de réussite possible, de distinguer un client fidèle d'un infidèle ? (cf graphique de droite).

La deuxième méthode est celle de l'apprentissage non-supervisé. Reprenons notre exemple et imaginons que Cornélius soit très timide et ne connaisse pas bien Zira. Il est donc amené à adopter une posture d'explorateur. Il va ainsi manipuler l'ensemble de ses brouillons afin de découvrir et utiliser par lui-même l'expertise qui se trouve implicitement dans son corpus de données. Par ce processus d'extraction des connaissances (i.e. : clustering), Cornélius finit par dessiner le graphique du centre sans faire appel à l'expertise de Zira.

La troisième méthode est celle de l'apprentissage semi-supervisé. Imaginons que ce soit cette fois-ci Zira qui soit timide. Ne connaissant pas bien Cornélius, elle décide de ne lui faire partager qu'une partie de son expertise en lui soufflant la moitié des réponses correctes. Cornélius est ainsi mené à articuler les méthodes d'apprentissage supervisé et non-supervisé afin de pouvoir réaliser le graphique du centre.

C'est donc par le biais de ces trois méthodes d'apprentissage qu'un agent cognitif artificiel peut accroître son intelligence, et donc, peut devenir plus performant et compétent dans l'accomplissement d'une tâche bien définie. Toute la complexité de l'apprentissage réside alors dans la bonne articulation du couple précision/généralisation afin d'éviter l'apprentissage de détails sans importance (i.e. : sur-apprentissage ; cf la courbe du graphique de droite) et les erreurs de catégorisation (i.e. : sur-généralisation ; cf la droite du graphique de droite).

Maintenant que nous avons dégagé les principes de base qui permettent aux machines d'apprendre à quantifier/qualifier le réel, nous proposons de revenir sur les deux paradigmes cognitifs qui sont aux fondements de l'intelligence artificielle. Ceci, dans l'objectif de mieux comprendre l'ensemble des conventions qui sous-tendent les différentes technologies d'apprentissage artificiel, les principes métrologiques qui leur sont sous-jacents, et ainsi de mieux saisir leurs implications pour la construction des connaissances (Desrosières, 2008).

3.2. Du modèle cognitiviste...

Commençons par présenter le paradigme cognitiviste. Notons d'emblée que celui-ci a pour objectif de comprendre le fonctionnement du cerveau à travers le modèle de l'ordinateur (Fodor, 1975). L'hypothèse fondamentale du cognitivisme peut donc être exposée comme suit :

> *« l'intelligence ressemble tellement à la computation dans ses caractéristiques essentielles que la cognition peut en fait se définir par des computations sur des représentations symboliques […]. Une computation est une opération effectuée ou accomplie sur des symboles, c'est-à-dire sur des éléments qui représentent ce dont ils tiennent lieu »* (Varela *et al.*, 1993, p. 73).

À l'instar de la psychologie du raisonnement, le cognitivisme considère que chaque représentation symbolique est définie par un ensemble de règles (i.e. : une syntaxe) et de significations (i.e. : une sémantique). Par exemple, pour Cornélius le cognitiviste, un état du monde (e.g. : le client fait un achat) est désigné par une représentation symbolique (e.g. : l'étiquette « nombre d'achats ») qui vient s'incarner dans un ensemble d'éléments physiques (i.e. : des bits) qui lui confèrent alors des qualités formelles lui permettant d'interagir avec d'autres représentations symboliques (e.g. : « nombre d'achats » peut être corrélé à « nombre de visites » ; le fait que le client fait un achat peut être associé au fait qu'il est venu plusieurs fois dans le magasin).

En synthèse, ici, un agent intelligent est un système de computation physique d'un ensemble de représentations symboliques (Newell, 1980). Et, ces représentations symboliques sont d'abord considérées du point de vue du code syntaxique puisque leurs sémantiques constituent des unités permanentes dont l'agent cognitif n'a pas besoin de se préoccuper (Haugeland, 1989). Par exemple, l'étiquette « nombre d'achats » de Cornélius est posée *a priori*. Durant son apprentissage, Cornélius se moque de ce qu'est un achat. Il reconnaît sa forme d'un point de vue syntaxique, mais il ne le comprend pas (Searle, 1983). Pourtant, et Zira le sait bien, cette représentation symbolique qu'est l'« achat » renvoie en fait à une interprétation qui ne tient pas compte des achats en ligne.

Dès lors, la tâche de l'apprenant cognitiviste consiste d'abord à rechercher une ou plusieurs hypothèse(s) s'accordant à un ensemble de représentations symboliques ; c'est-à-dire, de données étiquetées. Ainsi, les systèmes d'apprentissages cognitivistes sont en quelque sorte dépendants d'un livre de règles (Descombes, 1995) qui entraîne une relation à la connaissance de type fonctionnelle et descendante (Livet, 1995). Ils sont donc bien adaptés aux calculs de haut niveau qui impliquent des activités de planification, de calcul stratégique, tactique, etc. Néanmoins, compte tenu du goulot d'étranglement von Neuman[7] (1958), les apprenants cognitivistes peuvent éprouver certaines difficultés à effectuer des calculs engageant un grand nombre d'opérations séquentielles. De plus, ces apprenants sont sensibles à la perte ou détérioration des représentations symboliques et des règles d'inférences. En conclusion, si les apprenants cognitivistes ont de bonnes capacités de prédiction et généralisation, ils sont peu flexibles et adaptatifs.

3.3. … au modèle connexionniste

A l'opposé du cognitivisme, le paradigme connexionniste cherche à reproduire artificiellement la nature du cerveau humain. Il est forgé autour de la critique suivante : le modèle cognitiviste, qui est trop éloigné des caractéristiques neurophysiologiques du cerveau, est, en définitive, une émulation des savoir-faire experts (Varela, 1988). Or, au fur et à mesure des avancées du cognitivisme, il apparaît que l'intelligence la plus profonde et fondamentale est celle du nouveau-né qui acquiert des connaissances au quotidien. L'hypothèse principale du connexionnisme peut donc être présentée de la façon suivante : la cognition est le produit d'une masse d'entités élémentaires et non-intelligentes, qui, une fois associés les unes aux autres selon des règles implicites, s'auto-organisent pour exprimer des propriétés globales intéressantes (Rumelheart et McClelland, 1986).

Aussi, un des principes souvent présenté comme au cœur du fonctionnement des réseaux de neurones est la règle de Hebb : si deux neurones se stimulent ensemble, c'est quelles sont en train de se lier l'une à l'autre (Hebb, 1949). Par exemple, si une série de clients, sous la forme de

---

7. Cette expression permet de désigner les limites fonctionnelles de l'architecture von Neuman (1958) qui est actuellement le modèle de conception informatique le plus répandu. Plus concrètement, la largeur de bande qui permet aux informations de circuler entre l'unité de traitement et la mémoire est très réduite en rapport à la capacité de stockage et de calcul. C'est pourquoi, les agents von Neuman éprouvent, par exemple, des difficultés à effectuer des traitements minimaux sur de grandes quantités de données.

caractéristiques comportementales, est présenté à Cornélius le connexionniste, celui-ci va reconfigurer son réseau de neurones après chaque exposition en activant et inhibant ses perceptrons[8] jusqu'à temps qu'il puisse généraliser et reconnaître un profil de client (Rosenblatt, 1958). En associant à chaque client un état unique et global de son réseau de neurones, Cornélius exprime alors des configurations internes qui représentent les objets appris (Varela, 1988). C'est ainsi qu'il peut, par exemple, apprendre à associer et dissocier les caractéristiques comportementales des clients fidèles et infidèles sans faire appel à l'expertise de Zira[9].

En conséquence, la tâche de l'apprenant connexionniste consiste à s'auto-adapter à de nouvelles données plus ou moins structurées de façon à modifier ou préserver une configuration de neurones ; c'est-à-dire, une représentation distribuée, ou si l'on préfère, une représentation sous-symbolique (Smolensky, 1986). En ce sens, les systèmes d'apprentissages connexionnistes ne sont pas dépendants d'un ensemble de catégories pertinentes puisqu'« ils font « émerger » ces classifications du fonctionnement de leurs unités, reliées par des connexions dont les poids changent au cours de l'apprentissage » (Livet, 1995, p. 176). Ces systèmes impliquent ainsi une relation d'apprentissage qui est, cette fois-ci, de type interactionnelle et ascendante. Alors, les technologies d'apprentissage artificiel issues du paradigme connexionniste sont bien adaptées aux problèmes de bas niveau que sont la reconnaissance ou la classification en temps réduit. De plus, compte tenu de l'architecture distribuée du réseau de neurones, les apprenants connexionnistes savent effectuer rapidement des calculs qui impliquent de nombreuses opérations séquentielles. Ils peuvent également calculer des flux de données de façon à faire des anticipations de séquences continuent et instantanées (Hawkins et Blackeslee, 2004). Néanmoins, bien que ces apprenants soient flexibles et adaptatifs, ils ont des

[8]8. Un perceptron est une unité de traitement élémentaire d'un réseau connexionniste. Sa conception a été inspirée par la théorie cognitive de Friedrich A. Hayek (1952) et celle de Donald Hebb (1949). Le perceptron ne peut réaliser que des opérations très simples et est souvent appelé neurone formel en raison de sa similitude grossière avec le neurone du cerveau. Il est généralement caractérisé par une fonction de sortie (i.e. : fonction d'activation) qui permet de calculer pour chaque neurone d'entrée une valeur de sortie en fonction de son état d'activation.

[9]9. Ajoutons qu'il existe aujourd'hui des techniques statistiques qui permettent de réaliser ce même type d'apprentissage sans pour autant avoir recours aux réseaux de neurones. C'est le cas, par exemple, de l'algorithme des k-plus-proches-voisins. Aussi, bien que celui-ci soit très différent, sur le plan technique, des réseaux de neurones, il n'en reste pas moins implicitement influencé par le modèle connexionniste puisqu'il a été conçu dans le but d'accéder à une connaissance de type sous-symbolique (cf Smolensky, 1986).

capacités de systématicité et de productivité relativement limitées (Livet, 1995). La systématicité et la productivité sont pourtant au cœur des activités de calcul de haut niveau (Fodor et Pylyshyn, 1988).

3.4. Un processus sociocognitif en révolution

D'un point de vue marchand, l'enjeu du processus de documentation big data est de pouvoir assembler de façon cohérente l'ensemble du flux de données produites par les consommateurs. Celles-ci sont fabriquées à travers :

- les outils de traçage (sites web fréquentés, pages visitées, mots clés recherchés, etc.) ;
- les réseaux sociaux (qui parle de quoi ? Comment ? Avec qui ? Etc.) ;
- les technologies de géolocalisation (adresse Internet Protocol, Global Positioning System, etc.) ;
- et, les cartes de fidélité (fréquence de courses, panier moyen, type de produits achetés, etc.).

De cet assemblage de données doit alors découler un ensemble d'informations et documentations qui composent, en quelque sorte, « un programme d'évaluation des multiples actions ayant touché le client avant qu'il achète »[10].

En outre, nous l'avons déjà signalé, à la différence des données qui constituent les bases statiques que les technologies classiques viennent fouiller, les big data sont :

> *« trop volumineuses et arrivent trop rapidement pour être rangées dans des structures prédéfinies : la structuration plus faible fait que les corrélations entre données, et une visualisation perceptible à un œil humain sont impossibles à représenter du fait des quantités de lignes, graphes ou autres »*[11].

Compte tenu des capacités d'adaptation et de distribution des technologies d'apprentissage connexionniste, de nombreuses entreprises de stockage et traitement de données proposent alors des outils statistiques, logiciels et matériels largement influencés par ce modèle. C'est le cas, à différents degrés, des sociétés comme Statistical Analysis System et de sa solution Enterprise Miners Neural Network, Numenta et de sa solution Grok, ou encore

---

[10]10. Directeur marketing, extrait de magazine spécialisé dans les outils d'aide à la décision, janvier 2013.

[11]11. Architecte de stockage, extrait de magazine spécialisé dans le big data, avril 2013.

International Business Machines et de son projet Systems of Neuromorphic Adaptive Plastic Scalable Electronics.

Prenons l'exemple de la solution Grok qui est une des plus innovantes. En simulant le fonctionnement du néocortex humain, le HTM Cortical Learning Algorithm de Numenta (2011) est capable d'imiter les mécanismes neuronaux qui sont engagés dans la construction d'une (méta-)représentation. Plus clairement, par le biais d'une mémoire artificielle hiérarchique et séquentielle, Grok peut interpréter différentes séquences de comportements passés afin de prédire, à travers un ensemble de règles d'associations implicites, celles à venir. Par exemple, à partir des outils de traçage et des technologies de géolocalisation, Grok peut découvrir différents enchaînements d'actions et les combiner pour anticiper une séquence d'activité d'un consommateur bien identifié. Grok peut alors réviser ses prédictions de façon dynamique ; c'est-à-dire, au fur et à mesure de l'évolution de la situation de consommation observée. En synthèse, plus Grok se nourrit de cette nouvelle matière première que forment les big data, plus il découvre, apprend et anticipe de nouvelles représentations sous-symboliques et de nouvelles formes d'associations de ces représentations.

Par conséquent, du point de vue du processus de documentation que recouvre le big data marchand, l'application des technologies d'apprentissage non-supervisé et d'influence connexionniste conduit à instaurer une nouvelle forme de gestion ambidextre des données (Duncan, 1976). Au sens où ces technologies doivent permettre aux machines d'exploiter, de façon quasi-autonome, des logiques d'exploration d'un flux de données plus ou moins structurées. Autrement dit, à travers l'apprentissage de nouvelles méthodes d'extraction de connaissances, le processus de documentation big data doit permettre l'« expérimentation de nouvelles alternatives » (March, 1991, p. 85). Dans le même temps, en assemblant ces connaissances sous la forme de prédictions et de classifications en temps réel, il doit également permettre le « perfectionnement et l'extension des compétences » (March, 1991, p. 85). De ce fait, en instituant de nouveaux liens entre les logiques d'exploration et d'exploitation des données, le processus de documentation big data constitue, du point de vue marchand, une révolution sociocognitive (Freeman, 1994). Pour autant, cette révolution n'est pas dépourvue de toute contradiction : de façon traditionnelles, l'exploration a des « effets [qui] sont incertains, à long terme et souvent négatifs » alors que l'exploitation a des « effets [qui] sont positifs, rapides et prévisibles » (March, 1991, p. 85).

4. Abduction et organisation

C'est pourquoi, en guise de discussion, nous souhaitons maintenant pointer quelques implications organisationnelles de cette révolution.

Car, au regard des retours d'expérience issus de notre corpus, nous avons pu repérer que, pour lever l'antagonisme exploration/exploitation, les usagers du big data sont amenés à développer des logiques inférentielles qui sont de l'ordre de l'abduction[12] (Peirce, 1974). En effet, le processus de documentation big data permet d'assembler des données qui ont été, jusque-là, quasi-inexplorées. Du point de vue du récepteur, il est donc difficile de déduire les conséquences des phénomènes représentés par ces assemblages, puisqu'il n'existe pas vraiment de connaissances préétablies sur le sujet. De plus, le processus de documentation big data permet de combiner, n'ont pas un stock, mais un flux de données qui évoluent rapidement. Le récepteur peut donc éprouver certaines difficultés à induire les conséquences générales des phénomènes représentés par ces combinaisons, puisqu'elles renvoient à un ensemble de données toujours renouvelées. *A contrario*, le caractère abondant, fluide et inexploré des big data encourage le récepteur à rétrodéduire les règles permettant de saisir les corrélations qui assemblent ces données. C'est en ce sens que l'inférence abductive permet aux usagers du big data d'établir un compromis entre les exigences de performance que sous-tend la logique d'exploitation des données et celles d'expérimentation que sous-tend la logique d'exploration données.

Prenons l'exemple du professeur Mark Hansen[13]. Afin d'apprendre à ses étudiants le pouvoir de narration des données, ce dernier leur propose de placer des capteurs dans un ascenseur et un escalier de l'Université de New York. Il constate alors, non sans surprise et satisfaction, que les données ont bel et bien parlé : les étudiants utilisent l'ascenseur le matin plutôt que le soir. Il en rétrodéduit que les étudiants développent cette pratique, parce que, le matin, ils se sentent encore fatigués, alors que le soir, ils sont plutôt excités. Si nous avions participé au cours du professeur Hansen, nous aurions pu lui soumettre une autre hypothèse : si les étudiants prennent l'ascenseur le matin, c'est parce qu'ils s'y rendent individuellement et qu'il est ainsi facile d'y accéder ; ce qui n'est pas le cas du soir, où ils sont toujours groupés. Bien d'autres explications

---

[12]12. C'est-à-dire, un mode de raisonnement qui consiste à remonter du conséquent à l'antécédent à partir d'une règle causale plus ou moins inventée. Ainsi, l'abduction constitue, en quelque sorte, une des conséquences de la mise en pertinence de corrélation sans cause que peuvent parfois instaurer les technologies d'apprentissage artificiel (Rouvroy et Berns, 2010).

[13]13. Extrait des archives du *New York Times*, février 2013.

pourraient alors être explorées. Et, il émergerait ainsi une discussion sur ce qui doit être pris en compte pour comprendre ces données.

De la sorte, bien que cet exemple élémentaire porte moins sur les usages marchands du big data que sur les manières dont il est enseigné, il permet de montrer comment ce processus de documentation, en favorisant les raisonnements abductifs, peut encourager les situations de dissonance cognitive. C'est pourquoi, le processus de documentation big data constitue un outil intéressant pour organiser les situations de perplexité ; c'est-à-dire, des situations de remise en cause qui consistent moins à confronter des valeurs conventionnellement établies (Boltanski et Thévenot, 1991) qu'à rechercher à définir ce qu'est la valeur d'une chose[14] (qu'elle soit d'ordre économique, sociologique, politique, culturelle, technique ou encore scientifique ; Stark, 2009).

Conclusion

Reprenons. Le big data compose un processus de documentation (section 1). Et, si ce processus renvoie d'abord à un mouvement d'évolution sociotechnique (section 2), d'un point de vue sociocognitif et marchand, il constitue plutôt une révolution (section 3). Car, bien que les techniques d'assemblage des big data s'apparentent aux technologies classiques de fouille des données, elles recouvrent un procès de construction des connaissances qui est bien singulier. Dans les cas les plus avancés, le processus de documentation big data peut en effet transformer, de façon dynamique, ascendante et instantanée, les données produites par les consommateurs. Ceci, afin d'élaborer des classifications et prédictions qui ont pour finalité de documenter les acteurs du marché en temps réel.

Ainsi, à travers les technologies d'apprentissage artificiel non-supervisé et d'influence connexionniste, le processus de documentation big data instaure progressivement de nouvelles connexions entre les logiques d'exploitation et d'exploration. Les conséquences organisationnelles de ce changement sont remarquables puisque le succès d'une innovation, et plus largement d'une entreprise dans le long terme, repose sur la bonne articulation de ces deux

[14]14. Plus précisément, selon David Stark (2009), les situations de perplexité relèvent moins d'une mise en confrontation des cités de Luc Boltanski et Laurent Thévenot (1991) que d'une remise en cause des principes d'équivalence qui y sont associés. En d'autres termes, pour David Stark, il existe une sorte de cité de la perplexité où le grand est celui qui sait mettre à l'épreuve la notion même de grandeur ; c'est-à-dire, qui sait créer et/ou traiter des situations de réelle indétermination.

logiques (Mothe et Brion, 2008). Cependant, le processus de documentation big data renvoie, de fait, à une contradiction importante : comment articuler les incertitudes inhérentes aux logiques d'expérimentation et les exigences d'efficacité associées aux logiques de performance ? Pour lever cette contradiction, les usagers du système d'information big data sont conduits à expérimenter des raisonnements abductifs afin de signifier les performances de classification et prédiction des machines. De ce fait, en favorisant l'émergence de logiques rétroductives, le processus de documentation big data est susceptible d'encourager les situations de perplexité (section 4).

En conclusion, le processus de documentation big data peut permettre de développer des formes d'organisations marchandes plus hétérarchiques ; c'est-à-dire, qui sont capables de s'adapter efficacement à des environnements incertains et changeants (Stark, 2009). Cependant, il ne faut pas oublier qu'il peut également engendrer des effets de connaissances largement biaisés. Car, lorsque le professeur Mark Hansen raconte au gardien ce qu'il a trouvé, ce dernier lui répond en rigolant que, si les étudiants n'ont pas utilisé l'ascenseur le soir, ce n'est pas tellement parce qu'ils sont plus excités que le matin mais surtout parce que l'ascenseur a très mal fonctionné durant les soirées de la semaine de test ! Ainsi, si Mark Hansen avait travaillé pour le service marketing d'une grande société d'installation d'ascenseur, son abduction aurait pu instaurer, non plus une dissonance, mais une distorsion importante dans la relation client ; soit, plus largement, dans l'organisation du marché.

Références

Akrich M. (1989). La construction d'un système socio-technique. Esquisse pour une anthropologie des techniques. *Anthropologie et Sociétés*, vol. 12, n° 2, p. 31-54.

Barthes R. (1964). Rhétorique de l'image. *Communications*, vol. 4, n° 4, p. 40-51.

Bourdieu P. (1996). *Sur la télévision*, Raison d'agir, Paris.

Bryant R. E., Katz R. H., Lawoska E. D., (2008). *Big Data Computing : Creating revolutionary breakthroughs in commerce, science, and society*, http://www.cra.org/ccc/files/docs/init/Big_Data.pdf.

Brown B., Chui M., Manyika J., (2011). *Are you ready for the era of « big data »*. Report of Mc Kinsey Global Institute, October.

Bughin J., Chui M., Manyika J., (2010). *Clouds, big datas and smart assets : Ten tech-enabled business trends to watch*. Report of Mc Kinsey Global Institute, August.

Boltanski L., Thévenot L., (1991). *De la justification. Les économies de la grandeur*, Gallimard, Paris.

Callon M. (1986). Éléments pour une sociologie de la traduction : la domestication des coquilles St-Jacques et des marins pêcheurs dans la baie de St-Brieuc. *L'Année Sociologique*, n° 36, p. 169-208.

Chomsky N. (1965). *Aspects of the theory of syntax*, MIT Press, Cambridge.

Cochoy F. (2004). Introduction. La captation des publics entre dispositifs et dispositions, ou le petit chaperon rouge revisité. *La captation des publics*, Toulouse, PUM, p. 11-68.

Cornuéjols A., Miclet L., (2013). *Apprentissage artificiel. Concepts et algorithmes*, Eyrolles, Paris.

Courbet D. (2006). Les applications en sciences humaines à la publicité : de la psychanalyse à la socio-cognition implicite et au neuromarketing. *Humanisme et entreprise*, n° 276, p. 1-20.

Descombes V. (1995). *La denrée mentale*, Éditions de Minuit, Paris.

Desrosières A. (2008). *L'argument statistique. Pour une sociologie historique de la quantification*, PEM, Paris.

Dichter E. (1961). *La stratégie du désir*, Fayard, Paris.

Diebold F. X. (2003). « Big Data » : Dynamic Factor Models for Macroeconomic Measurement and Forecasting. *Advances in Economics and Econometrics*, Cambridge, CUP, p. 115-122.

Duncan R. B. (1976). The ambidextrous organization : designing dual structures for innovation. *The management of organization*, New York, North-Holland, p.167-188.

Fodor J. A. (1975). *The Language of Thought*, HUP, Cambridge.

Fodor J. A., Pylyshyn Z. (1988). Connectionism and cognitive architecture : A critical analysis. *Cognition*, n° 28, p. 3-71.

Freeman C., (1994). The economics of technical change. *Cambridge Journal of Economics*, vol. 18, n° 5, p. 436-514.

Gomez P.-Y. (2006). « Information et conventions ». Le cadre du modèle général. *Revue française de gestion*, vol. 1, n° 160, p. 217-240.

Haugeland J. (1989). *L'Esprit dans la machine : Fondements de l'intelligence artificielle*, Odile Jacob, Paris.

Hawkins J., Blakeslee S., (2004). *On Intelligence*, Times Books, New York.

Hayek F. A. (1952). *The Sensory Order: An Inquiry Into the Foundations of Theoretical Psychology*, UCP, Chicago.

Hebb D. O. (1949). *The Organization of Behavior : A Neuropsychological Theory*, Wiley, New York.

Hughes T. (1987). The Evolution of Large Technical Systems. *The Social Construction of Technological Systems*, Cambridge, MIT Press, p. 51-82.

Kessous E. (2012). *L'attention au monde. Sociologie des données personnelles à l'ère numérique*, Armand Colin, Paris.

Latour B. (1993). Le « pédofil » de Boa Vista ou la référence scientifique. *La clef de Berlin et autres leçons d'un amateur de science*, Paris, La Découverte.

Laval C. (2007). *L'homme économique : Essai sur les racines du néolibéralisme*, Gallimard, Paris.

Livet P. (1995). Connexionnisme et fonctionnalisme. *Intellectica*, vol. 2, n° 21, p. 175-197.

Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Hung Byers A., (2011). *Big data : The next frontier for innovation, competition and productivity*. Report of Mc Kinsey Global Institute, June.

March J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, vol. 2, n° 1, p. 71-87.

Mashey J. (1998). *Big Data... and the Next Wave of InfraStress*, https://www.usenix.org/legacy/event/usenix99/invited_talks/mashey.pdf.

Mothe C., Brion S., (2008). Innovation : exploiter ou explorer ? *Revue Française de Gestion*, n° 187, p. 101-108.

Neumann (von) J. (1958). *The Computer and the Brain*, YUP, London.

Newell A. (1980). Physical symbol systems. *Cognitive Science*, vol. 4, n° 2, p. 135–83.

Numenta (2011). *Hierarchical Temporal Memory including HTM Cortical Learning Algorithms*, http://numenta.org/resources/HTM_CorticalLearningAlgorithms.pdf.

Peirce C. S. (1974). *Collected Papers*, HUP, Cambridge

Rosenblatt F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, vol. 65, n° 6, p. 386-408.

Rouvroy A., Berns T. (2010). Le nouveau pouvoir statistique. Ou quand le contrôle s'exerce sur un réel normé, docile et sans événement car constitué de corps numériques. *Multitudes*, n° 40, p. 88-103.

Rumelhart D. E., McClelland J. L., (1986). *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*, MIT Press, Cambridge.

Simon H. A. (1982). *De la rationalité substantive à la rationalité procédurale*, http://www.mcxapc.org/fileadmin/docs/lesintrouvables/simon5.pdf.

Smolensky P. (1986). On the Proper Treatment of Connectionism. *Behavior and Brain Sciences*, n° 11, p. 1-74.

Searle J R. (1983). Minds, brains and programs. *The Minds I : fantasies and reflections on self and soul*, Hardmondsworth, Penguin Books, p. 353-373.

Stark D. (2009). *The Sense of Dissonance. Accounts of Worth in Economic Life*, PUP, Princeton.

Tversky A., Kahneman D., (1974). Judgement under Uncertainty : Heuristics and Biases. *Science*, vol. 185, n° 4157, p. 1124-1131.

Varela F. (1988). *Cognitive science : a cartography of current ideas*, MIT Press, Cambridge.

Varela F., Thompson E., Rosch E., (1993). *L'inscription corporelle de l'esprit*, Seuil, Paris.

Weil-Barais A. (2005). *L'homme cognitif*, PUF, Paris.

HANDLING THE DATAS. DOCUMENTING THE MARKET

Organizational implications of the big data movement

JEAN-SÉBASTIEN VAYRE

What is big data? How to characterize it? What is its impact on the organization of the market? From a merchant point of view, big data consist in transforming the traces of consumers' activities into information which are transmit to the market participants. Big data is, therefore, a process of documentation. First, we argue that the dynamics of this process refers to a socio-technical evolution and a socio-cognitive revolution. Then, we underline the implications of this revolution on the market organization. In conclusion, we note that if big data documentation process can promote the reactivity and adaptability of commercial organizations, it can also lead to important knowledge biases in socioeconomic terms.